Things in data engineer – 1 – 115+hrs

Core java

Hadoop Mapreduce framework

Hive

Scoop

Hbase/Cassendra

Maven

Junix

Java code coverage(JACOCO)

Sonar Reporting

// Fling

// MongoDB

DE-2 – 115+ hrs

Scala

Spark

Kafka

Elastic search

Role:

Client based solutions

Mapreduce is the heart of Data Engineering. Distributed codes is imp.

Big Data at a high level is problem to solve:

1. Huge data(Volume)
2. Large processing time

Distributed codes are the base of big data

More the nodes, less the time taken

What is big data and hadoop?

Big data != Hadoop

C,C++, Java,Scala,Python

MySQL, DB2, Postrge

WebFramework: PHP, Angular,ReactJS

(cateogries)

Big data: hadoop, scala,spark, Hydin, CPCC

Hadoop is popular since it is open source. Written in Java primarily.

Speed with Java Hadoop cannot be compared to python hadoop.

Hadoop

1. Big data framework in Java
2. Launched in 2003 // Check
3. Distributed
4. Used since 2010
5. Has two components:
   1. Map reduce – to process the data
   2. HDFS – to store data –
      1. Hadoop cluster: network of computers, connected together by network,
      2. Capactity: add the specification of each component
      3. How does this happen: HDFS breaks the data into blocks, of maximum size 128MB (by default, can be changed/configured). No block can have greater than that. The blocks then can be stored on any node.
      4. Hadoop 1,2,3 is there in the machine, 2 is used, 3 is coming up.
      5. What if a node crashes? Lose of data? No. Data is duplicated onto another node, so that data is not lost.
      6. Wasting space because of this? But storage is cheap, data must not be lost. This also happens in cloud, So that data is not lost
      7. Do not store large amount of very small files. Affects name node, which keeps the details of files – names, replication factor, blocks, etc.
      8. At a high level : We need Java API to write or use mapreduce program, packed in .jar file,
      9. Jar file is then sent to hadoop cluster. Based on the resources available, the codes are processed on different nodes. And the results are integrated.
      10. Hadoop ensures minimal data traversal happens
      11. DATA DOES NOT TRAVEL.
   3. MapReduce – Mapper Phase, Reducer place
      1. Hive – Table format data
      2. Hbase: array/json type of data.
      3. Oozie/Airflow; Schedulers – used to automate jobs

What to do?

Linkdin, Naukri – search for jobs, make a list of requirements, which companies are looking for what.

Python is older than Java. But java was used for hadoop, because of Java – robustness feature.

Apache – Open source usually

Cloudera, databricks, MapR, - Modified Apache products, and selling with support.